LIDet: Language-Guided Iterative Object Detection

Anonymous EMNLP submission

Abstract

This paper proposes LIDet, a language-guided iterative object detection framework, designed to address challenges in open-vocabulary object detection, such as missed detections of small objects and rare categories, as well as false positives. Without retraining the detection model, the method constructs a four-stage closed-loop process:"image preprocessing → multimodal perception → object detection → language reasoning." Leveraging the semantic reasoning capabilities of large language models (LLMs), LIDet generates potential missing object categories and their spatial relationships based on detected objects and scene descriptions. This guides the visual detector to dynamically crop and re-examine image regions. Experiments demonstrate that LIDet achieves an average improvement of 3% in Acc@IoU=0.25 on the RefCOCO series datasets compared to the MQADet and outperforms the original detection model. Although computationally intensive, LIDet establishes a language-vision interaction mechanism at the semantic level, offering a novel approach to multimodal reasoning and open-vocabulary object detection.

1 Introduction

Open-domain object detection aims to overcome traditional closed-set limitations by dynamically recognizing unknown objects. Current research focuses on visual feature extraction and vision-language alignment.

The former leverages adaptive strategies, with AdaZoom (Xu et al., 2022) employing a multi-scale approach and ZIO (Pang et al., 2022) utilizing multi-resolution processing, to improve the detection of small objects, but both lacks deep semantic understanding. The latter, particularly CLIP-based (Radford et al., 2021) methods, incorporates textual matching but remains vision-dominated without utilizing language models' reasoning capabilities.

Notable performance drops occur with fine-grained small objects and rare categories, due to resolution limitations and semantic ambiguity in current convolutional or Transformer-based (Vaswani et al., 2017) networks.

We propose integrating LLMs' semantic reasoning into detection. Current LLM applications like LLMDet (Fu et al., 2025) and MQADet (Li et al., 2025) only generate pseudo-labels or filter results, lacking textual feedback in the detection pipeline.

To enable multi-round vision-language interaction during detection by leveraging LLMs' reasoning capabilities, we propose **LIDet**, a language-guided iterative object detection framework. Without retraining the detection model, LIDet guides multi-round detection across different regions using LLMs' semantic reasoning.

The framework consists of four stages: the **image preprocessing** stage performs cropping, scaling, and super-resolution; the **multimodal perception** stage generates image descriptions and parses potential targets; the **detection model inference** stage identifies objects based on text prompts; and the **language model reasoning** stage infers focus regions for potentially missed objects. The key innovation lies in using LLMs to deduce relative object positions and guiding the detector to re-examine these regions, establishing a closed-loop vision-text interaction.

In summary, Our main contributions are as follows:

- We propose **LIDet**, a **training-free iterative detection framework** that collaborates super-resolution, multimodal, and language models with detectors in a pipeline, significantly improving open-domain detection accuracy.

- Experiments show LIDet outperforms **MQADet by 3%** and **baseline models by 19%** on the RefCOCO datasets, demonstrating its effectiveness for fine-grained detection.

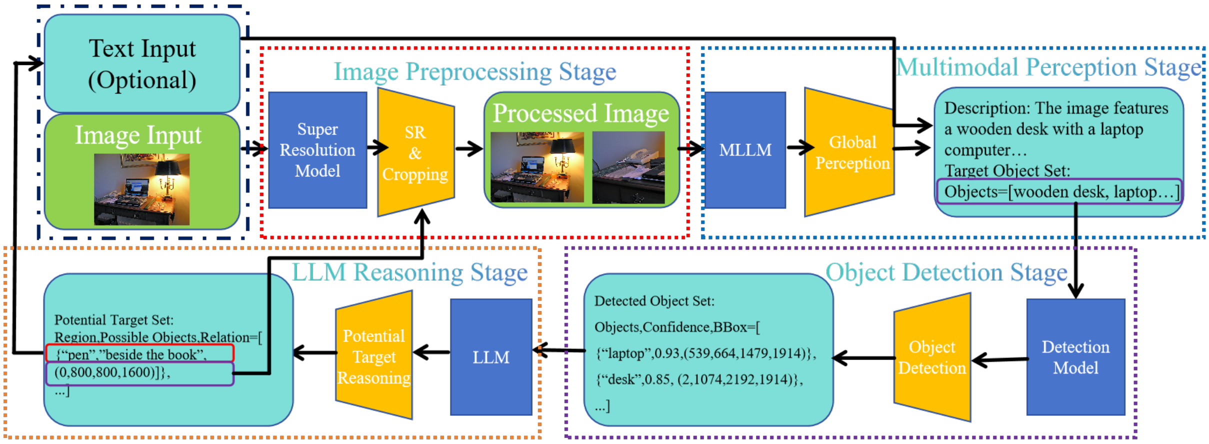

Figure 1: An overview of LIDet frame, including four stages: Image Preprocessing, Multimodal Perception, Object Detection and LLM Reasoning.

2 Related Work

2.1 Open-Vocabulary Detection

Open-vocabulary detection (OVD) extends beyond fixed categories by leveraging vision-language alignment. CLIP (Radford et al., 2021) enables zero-shot transfer and has inspired approaches such as ViLD (Gu et al., 2021) for detector distillation and RegionCLIP (Zhong et al., 2022) for region-level representation learning. Recent work further improves fusion efficiency (Liu et al., 2024; Cheng et al., 2024; Yao et al., 2024; Fu et al., 2025).These methods struggle with complex cross-modal reasoning via vision-language pretraining. A more promising approach is to use language models as reasoning agents to enhance visual detection with their strong inference capabilities.

2.2 Vision-Language Models

Modern vision-language systems build upon alignment foundations like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), evolving into interactive reasoning architectures.LLaVA (Liu et al., 2023) establishes visual-text coupling through projected embeddings and MLP fusion. Qwen-VL (Wang et al., 2024) extends this with dynamic resolution support for complex spatial tasks. Flamingo (Alayrac et al., 2022) innovates with visual token compression for efficient cross-modal attention. These models demonstrate robust visual perception capabilities, effectively translating images into textual representations. Their progress enables seamless integration of visual data with language-based reasoning frameworks, supporting downstream tasks in our work through unified multimodal understanding.

2.3 Super Resolution

Multimodal object detection accuracy relies on image quality, where Single Image Super-Resolution (SISR) enhances low-resolution inputs. Early interpolation and shallow CNN methods (Dong et al., 2015, 2016; Kim et al., 2016) caused over-smoothing, while GAN-based approaches (e.g.Ledig et al., 2017) improved texture generation via adversarial learning. Later, Wang et al. (2018) stabilized training with relative discriminators, and Real-ESRGAN (Wang et al., 2021) advanced real-world modeling. SwinIR's Transformer architecture (Liang et al., 2021) excelled in detail reconstruction using attention. Current research emphasizes multimodal fusion and degradation-aware designs, evolving from pixel-level to semantic-physical modeling.

3 Methodology

3.1 Image Preprocessing

Image quality (clarity and object size) critically impacts open-domain detection accuracy. Low-resolution images degrade small-object features and bounding box localization, as shown in SNIP (Singh and Davis, 2018). While AdaZoom (Xu et al., 2022) and Zoom-In&Out (Pang et al., 2022) enable region magnification, their vision-based selection introduces contextual noise.

We innovatively propose a language-guided crop-and-zoom strategy. By leveraging multimodal scene descriptions and semantic reasoning of detected objects, we precisely determine cropping regions that preserve effective contextual information while avoiding noise interference. Furthermore, we employ Real-ESRGAN (Wang et al., 2021) to com-

Method	RefCOCO			RefCOCO+			RefCOCOg	
	Val	testA	testB	Val	testA	testB	Val	test
G-DINO	48.21 / 42.85	49.83 / 45.08	40.50 / 36.58	49.66 / 41.56	50.58 / 43.98	43.51 / 37.51	40.76 / 38.18	41.96 / 39.24
MQADet (G-DINO)	66.59 / 60.47	64.01 / 60.03	67.20 / 61.70	57.29 / 49.50	55.07 / 48.51	56.87 / 50.18	66.10 / 61.45	67.91 / 62.90
LIDet (G-DINO)	69.23 / 59.68	68.47 / 58.46	66.36 / 60.81	64.37 / 52.49	61.48 / 50.92	62.82 / 53.62	70.32 / 58.61	66.47 / 57.39
MM-GDINO	50.21 / 44.37	51.87 / 46.53	41.40 / 38.02	51.29 / 42.47	50.63 / 44.21	44.20 / 39.14	41.20 / 39.43	42.13 / 39.76
LIDet (MM-GDINO)	71.43 / 59.57	70.15 / 60.47	70.63 / 59.48	65.28 / 54.82	63.04 / 51.48	62.49 / 53.73	71.17 / 58.93	67.54 / 58.47
Yolo-World	38.15 / 32.65	42.70 / 38.36	32.97 / 28.47	37.82 / 31.06	38.20 / 33.77	35.32 / 30.65	40.11 / 36.99	43.05 / 38.51
MQADet (Yolo-World)	62.79 / 56.81	60.59 / 55.28	62.13 / 55.65	56.97 / 48.31	52.91 / 46.88	55.47 / 48.84	62.50 / 57.55	65.57 / 60.44
LIDet (Yolo-World)	64.36 / 53.19	63.73 / 52.86	64.09 / 55.36	60.34 / 49.36	59.46 / 47.25	58.63 / 49.27	63.45 / 52.82	64.03 / 53.46

Table 1: Evaluation of the LIDet framework against various detection models across RefCOCO, RefCOCO+, and RefCOCOg datasets (with provided standard val/testA/testB splits), using **Acc@0.25** and **Acc@0.5** as evaluation metrics in the form of Acc@0.25/Acc@0.5. The LIDet parameters are fixed at $k = 2$ (iterations), $m = 3$ (candidate targets per iteration), $\alpha = 1.5$ (image zoom ratio) using Qwen2.5-14B-Instruct for reasoning.

pensate for resolution loss during the magnification process.

$$I' = \text{Real-ESRGAN}(\text{Crop}(I, LLM(S, E)) \quad (1)$$

where S is the description of the image I, E is the detected object set.

3.2 Multimodal Perception

We generate scene descriptions S using pretrained multimodal models. For structured detection inputs, we employ prompt engineering and parse S with prompt P using instruction-tuned LLMs considering LLaVA's (Liu et al., 2023) limitations in structured output.

$$U_0 = \{O_j\}_{j=1}^{N} = \text{LLM}(P \oplus S) \quad (2)$$

Merging these target set U_0 with previous-round potential targets U_{new} yields the final target object set $U = U_0 \cup U_{\text{new}}$.

3.3 Object Detection

During the detection stage, the target set U is fed into the detection model to obtain the current round's detected objects E_0, which can be formally expressed as:

$$E_0 = \{(c_i, b_i) \mid c_i \in U, b_i = \text{Det}(I', c_i), \text{score}(b_i) > \tau\} \quad (3)$$

where c_i represents the detected object category, and b_i is the bounding box with confidence over threshold τ, U is the set of target classes, and $\text{Det}(\cdot)$ denotes the detection model. Subsequently, the newly detected objects E_0 are aggregated with the overall detected target set E through set union: $E = E_0 \cup E_{old}$.

3.4 LLM Reasoning

The model outputs both potential targets U_p and their spatial relationships relative to existing targets with inputs of annotated frames. Given these, we perform region localization and cropping based on reference bounding boxes $b_i \in E$ and spatial relationships:

$$Area_{crop} = \text{Crop}\,(I, \text{Scale}(b_i, \alpha)) \quad (4)$$

where α is the scaling ratio relative to the reference bounding box area $|b_i|$. This generates candidate regions for subsequent detection iterations.

The complete potential target set for the next iteration is then:

$$U_{new} = \bigcup_{m=1}^{M} \text{Top-m}(P(U_p|E)), \quad U = U_0 \cup U_{new} \quad (5)$$

where M controls the number of potential targets per iteration.

4 Experiments

4.1 Implementations

For the LIDet framework's four-stage pipeline, we conduct benchmark evaluations on the RefCOCO series datasets using the following open-source

Method	Val	testA	testB
MQADet	66.59 / 60.47	64.01 / 60.03	67.20 / 61.70
LIDet(α=1)	67.54 / 61.56	64.29 / 62.65	66.03 / 60.58
LIDet(α=1.5)	69.23 / 59.68	68.47 / 58.46	66.36 / 60.81
LIDet(α=2)	60.49 / 30.98	58.28 / 26.43	59.33 / 28.71

Table 2: Performance comparison between MQADet and LIDet with different zoom ratios α on the RefCOCO dataset (all based on GroundingDINO). Each cell reports Acc@0.25 / Acc@0.5. LIDet uses $k = 2$, $m = 3$ with Qwen2.5-14B-Instruct.

models in Appendix A. We adopt the Acc@IoU metric from MQADet for consistent performance comparison. The experiments are implemented with Python 3.10, PyTorch 2.1.2, and CUDA 12.1, running on a hardware platform equipped with 3× NVIDIA RTX 4090 GPUs.

4.2 Results

As Table 1 shows, our LIDet framework achieves superior **Acc@0.25** performance (+3% over MQADet on average, +7% on RefCOCO+) when configured with 2 iterative detection rounds and 3 potential targets. The significant improvement on RefCOCO+ stems from our model's reduced textual dependency and scene-based positional inference capability, which compensates for the dataset's prohibition of location words in referring expressions. However, we observe notable **Acc@0.5** degradation due to image preprocessing: detection boxes marked on zoomed sub-images ($\alpha\times$) then rescaled cause inherent IoU reduction, lowering the theoretical maximum from 1 to $1/\alpha$ even for perfect detection. This analysis is further validated by our zoom ratio ablation studies in subsection 4.3.

4.3 Ablation Study

Controlled zoom ratios. Under controlled conditions, we tested the G-DINO model with LIDet framework on RefCOCO with three zoom ratios (α) as shown in Table 2. At $\alpha = 1$, both Acc@0.25 and Acc@0.5 matched MQADet's performance. The Acc@0.25 for $\alpha = 1.5$ surpassed that of $\alpha = 1$, validating the effectiveness of zoom-in for regional focus, while $\alpha = 2$ degraded to the Acc@0.5 level of $\alpha = 1$, corroborating our IoU scaling analysis.

Size of LLM. As shown in Table 3, our investigation of language model scaling effects reveals that Acc@IoU remains remarkably stable across different model sizes. This indicates that the scene descriptions generated by multimodal perception models primarily require only fundamental reasoning abilities and commonsense object relationship understanding from the language model, rather than advanced linguistic capabilities.

Method	Val	testA	testB
LIDet(7B)	67.14 / 59.42	67.56 / 58.17	66.25 / 59.52
LIDet(14B)	69.23 / 59.68	68.47 / 58.46	66.36 / 60.81
LIDet(32B)	69.20 / 59.57	69.55 / 59.03	70.08 / 60.56

Table 3: Performance comparison between LIDet with different size of Qwen2.5-Instruct Model on the RefCOCO dataset (all based on GroundingDINO). Each cell reports Acc@0.25 / Acc@0.5. LIDet uses $k = 2$, $m = 3$, $\alpha = 1.5$.

Method	**Val**	**testA**	**testB**
LIDet(k=1, m=3)	65.40 / 57.57	62.91 / 55.83	64.82 / 58.49
LIDet(k=2, m=3)	69.23 / 59.68	68.47 / 58.46	66.36 / 60.81
LIDet(k=3, m=3)	71.06 / 60.12	69.74 / 59.84	67.35 / 60.87
LIDet(k=2, m=5)	69.35 / 59.82	68.52 / 58.60	66.37 / 60.92
LIDet(k=2, m=10)	69.55 / 59.74	69.61 / 58.15	67.51 / 59.78

Table 4: Comparison of LIDet performance with varying hyperparameters (k, m, $\alpha = 1.5$) on the RefCOCO dataset, evaluated using Acc@0.25 / Acc@0.5. All results are based on GroundingDINO.

Iteration rounds and candidate targets. Through iterative optimization, we reformulate detection as a search task. As Table 4 shows, hit probability grows with search space expansion due to accumulating contextual information from detected targets $\mathcal{D}_t = \{d_1, ..., d_t\}$. Formally, with scene description $\mathcal{S}$ and detection accuracy $P_{\text{detect}} \in [0, 1]$, larger $|\mathcal{D}_t|$ enhances reasoning by providing richer constraints for subsequent predictions.

5 Conclusion

In this paper, we propose a language-guided iterative object detection framework called LIDet, consisting of four main stages: image preprocessing stage, multimodal perception stage, detection model stage, and LLM reasoning stage. By establishing a closed-loop interaction mechanism between visual detection and language reasoning, our method achieves an average improvement of 3% in Acc@0.25 metrics on the RefCOCO series datasets compared to the MQADet framework, and an average 19% improvement over the baseline model. These results validate the effectiveness of the language model-guided iterative optimization strategy for open-vocabulary object detection. We hope this work will inspire future research in multimodal domains regarding image-text interactive reasoning.

Limitations

The experimental results demonstrate that the core value of the LIDet framework lies in its iterative vision-language interaction mechanism.However, as shown in the ablation studies 4.3, the current approach suffers from significant performance degradation by cropping zoomed regions and rescaling bounding boxes back to the original image in terms of Acc@0.5 metric. Future work could explore alternative strategies, such as center-based proportional scaling of bounding boxes in sub-images and propose more scientifically rigorous evaluation metrics.

The method's computational overhead constitutes another practical constraint. Benchmark tests reveal an average processing time of 40.1s per image, representing a 20-fold increase over baseline detectors like G-DINO (1.9s/image). This substantial latency originates from the language model's repeated autoregressive decoding cycles (minimum 4 passes per candidate region), with temporal complexity growing linearly with iteration count. Consequently, the current implementation fails to meet the throughput requirements of time-sensitive applications.

Ethics Statement

All models and datasets used in this work are publicly available, and their original development processes incorporated ethical reviews. Note that text generation inherently carries stochasticity, even with safety-aligned instruction fine-tuning (as adopted in prior works), there remains a non-zero probability of generating unexpected outputs. We may mitigate this by: lowering sampling temperature, and increasing confidence thresholds during decoding. Besides, we used Deepseek for grammar suggestions and writing refinement. All scientific content, experimental design, and analysis were conducted by the authors.

Acknowledgements

We thank Prof.Deng for his guidance and support on this research, also appreciate the helpful discussions with our senior lab members regarding experimental results analysis.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

Chao Dong, Chen Change Loy, and Xiaoou Tang. 2016. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. 2025. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. *arXiv preprint arXiv:2501.18954*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

Caixiong Li, Xiongwei Zhao, Jinhang Zhang, Xing Zhang, Qihao Sun, and Zhou Wu. 2025. Mqadet: A plug-and-play paradigm for enhancing open-vocabulary object detection via multimodal question answering. *arXiv preprint arXiv:2502.16486*.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. 2022. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2160–2170.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Bharat Singh and Larry S Davis. 2018. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

Jingtao Xu, Ya-Li Li, and Shengjin Wang. 2022. Adazoom: Towards scale-aware large scene object detection. *IEEE Transactions on Multimedia*, 25:4598–4609.

Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. 2024. Detclipv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27391–27401.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803.

Appendix

A Model Details

Table 5 displays the weight checkpoints of the adopted open-source models, where all weights were obtained from either Hugging Face's platform or the official GitHub repositories of the corresponding projects. All third-party models and tools used in this work are under open-source licenses.

Model	Checkpoints
Real-ESRGAN	RealESRGAN_x4plus.pth
LLaVA-v1.5	Liuhaotian/llava-v1.5-7b Openai/clip-vit-large-patch14-336
GDINO	groundingdino_swint_ogc.pth
YOLO-World	Yolo_world_v2_xl_obj365v1_goldg cc3mlite_pretrain.pth
MM-GDINO	grounding_dino_swin- t_pretrain_obj365_goldg_v3det_2023_ 1218_095741-e316e297.pth
Qwen2.5	Qwen/Qwen2.5-7/14/32B-Instruct

Table 5: Models and Checkpoints Used in Different Stages

B Samples of LIDet

Here in Figure 2 we present a visual comparison of the detection results produced by the GroundingDINO model before and after applying the LIDet framework, using representative sample images from the COCO dataset.

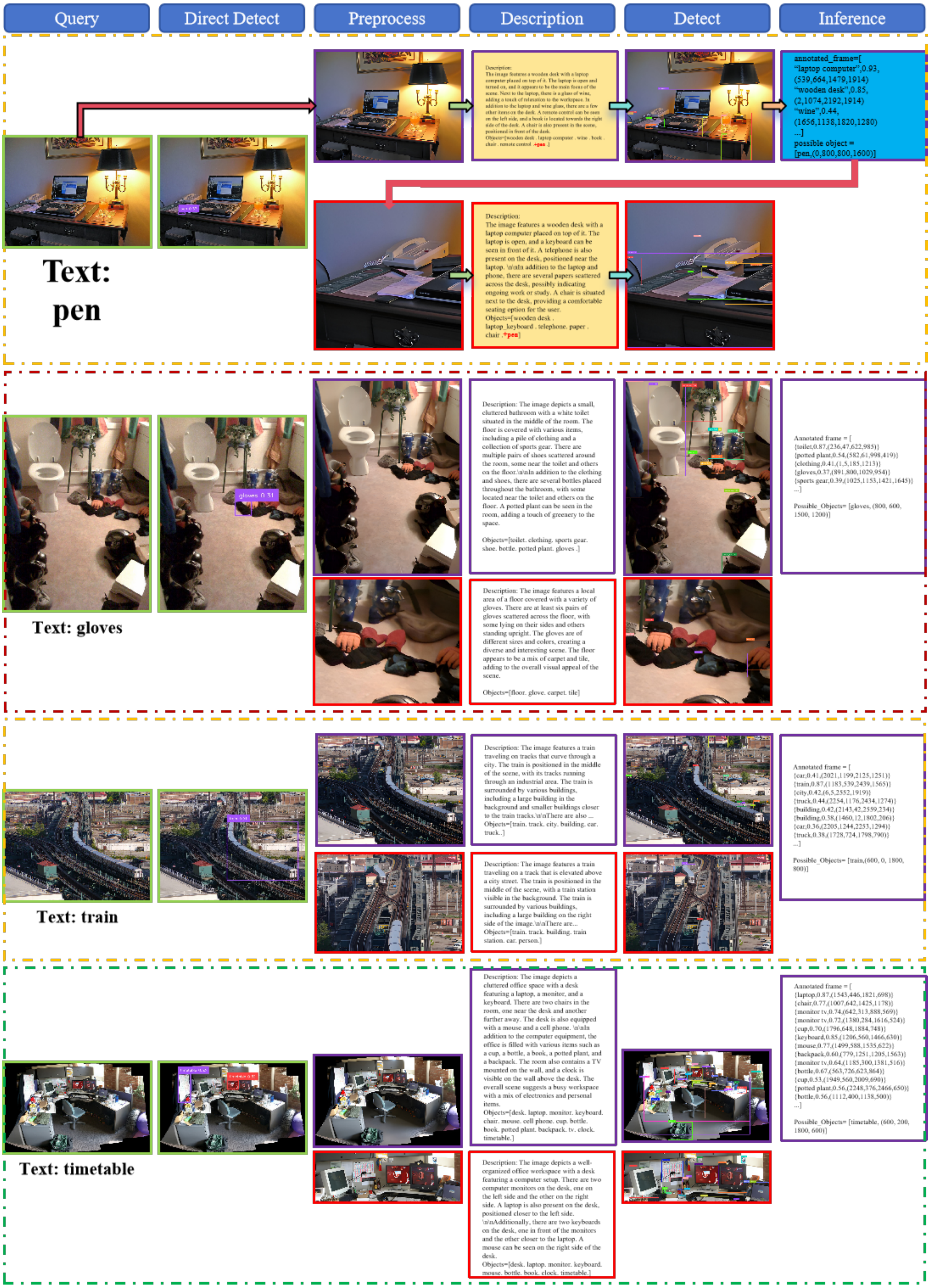

Figure 2: Some samples of our LIDet frame based on Groundingdino for detection.